
Some Theory For Practical Classifier Validation

Eric Bax* and Ya Le†

Abstract

We compare and contrast two approaches to validating a trained classifier while using all in-sample data for training. One is simultaneous validation over an organized set of hypotheses (SVOOSH), the well-known method that began with VC theory. The other is withhold and gap (WAG). WAG withholds a validation set, trains a holdout classifier on the remaining data, uses the validation data to validate that classifier, then adds the rate of disagreement between the holdout classifier and one trained using all in-sample data, which is an upper bound on the difference in error rates. We show that complex hypothesis classes and limited training data can make WAG a favorable alternative.

1 Introduction

One goal in machine learning is to use all available data for training and still compute effective test error bounds. A seminal result by Vapnik and Chervonenkis [1] showed that we can use the same data for training and validation, if we use simultaneous validation over all hypotheses in the hypothesis class used for training. Using this setup, the probability that test error rate for the trained classifier exceeds its training error rate by at least ϵ is at most

$$m(n)b(n, \epsilon),$$

where n is the number of training examples, $m(n)$ is the number of hypotheses in the hypothesis class, and $b(n, \epsilon)$ is an upper bound on a deviation of at least ϵ between empirical and actual means over n samples. This bound was revolutionary: since $m(n)$ grows as a polynomial in n for many hypothesis classes and $b(n, \epsilon)$ shrinks exponentially in n , we can be confident that if training selects a classifier with a low training error rate, it will have a low test error rate as well, with high probability, given sufficiently many training examples.

The original VC bound used shattered sets and VC dimension to bound m . There have been many improvements and variations on hypothesis counting, including the luckiness framework [2], margin bounds for hyperplanes [3], and PAC-Bayesian bounds [4]. These advances all build on the same basic concept of validation based on the number of hypotheses in a hypothesis class (or a distribution over those hypotheses), selected prior to examining the training data.

Similarly, there have been many improvements and variations on the Hoeffding bounds [5] originally used to bound b , including Azuma bounds [6], McDiarmid bounds [7], direct computation of bounds by binomial inversion [8, 9], bounds that take advantage of the low variance for accurate classifiers [10, 11], and other concentration inequalities [12]. These results focus on b , so they can be applied to a single classifier as easily as to a bound over multiple hypotheses.

Consider an alternative validation method: withhold $v < n$ validation examples. Train a *holdout classifier* on the remaining $n - v$ examples. Next, train a *full-data classifier* on all training examples. Then compute (in the transductive setting, where test inputs are known [3]) or validate the rate of disagreement Δ between the holdout and full-data classifiers over test data.

*baxhome@yahoo.com

†yle@stanford.edu

Since the holdout classifier is independent of the validation examples, we can use single-classifier validation over those examples to bound the test error rate of the holdout classifier. The full-data classifier test error rate is at most d greater than the holdout classifier test error rate. (In the worst case, every disagreement is an error for the full-data classifier.) So the probability that the full-data classifier test error rate exceeds the holdout classifier error rate over the validation data by at least ϵ is at most

$$b(v, \epsilon - \Delta)$$

in the transductive case. (We will concentrate on the transductive case in the main body of this note.) Call this the WAG (withhold and gap) bound.

Compare the WAG bound to the bound based on validation over a set of hypotheses (SVOOSH). The WAG bound has the advantage of not requiring simultaneous bounds over m hypotheses. However, it uses fewer examples for validation (v vs. n), and it has to include the rate of disagreement Δ as part of the bound range, ϵ .

In this note, we explore conditions for WAG to provide stronger error bounds than traditional methods. We also offer some intuition around the process of validation, considering for the traditional methods how many examples are needed to “select a hypothesis” vs. to “validate the selected hypothesis” and for WAG how similarity between a holdout and the full-data classifier can indicate that learning has been effective.

2 The Cost of Simultaneous Validations

For SVOOSH, consider how many extra examples are needed because we validate multiple hypotheses instead of just one. For SVOOSH,

$$\delta = m(n)b(n, \epsilon).$$

Using Hoeffding bounds [5],

$$b(n, \epsilon) = e^{-2n\epsilon^2}.$$

So

$$\delta = m(n)e^{-2n\epsilon^2}.$$

Let s be the value such that

$$e^{-2s\epsilon^2} = \frac{1}{m(n)}.$$

Then

$$\delta = m(n)e^{-2s\epsilon^2}e^{-2(n-s)\epsilon^2} = e^{-2(n-s)\epsilon^2} = b(n-s, \epsilon),$$

which is the probability of bound failure for single-classifier validation using $n-s$ examples.

Solve Equality 2 for s :

$$s = \frac{\ln m(n)}{2\epsilon^2}.$$

Since s is $O(\ln m)$, the number of extra examples required to validate multiple hypotheses grows very slowly in the number of hypotheses. Suppose $m(n) = n^d$, and call d the *dimension* of the hypothesis set – similar to VC dimension. Then s is $O(d \ln n)$. So s depends strongly on dimension d . Also, the portion of examples required because of multiple-hypothesis validation, $\frac{s}{n}$, shrinks quickly with the number of examples: it is $O(\frac{\ln n}{n})$.

3 A Limit on Validation Set Size for WAG

What does this have to do with WAG vs. SVOOSH? In SVOOSH, $\delta = b(n-s, \epsilon)$, meaning that the bound is the same as for using s examples to select a hypothesis from the set and then using the remaining $n-s$ examples to validate the selected hypothesis. In WAG, with $\delta = b(v, \epsilon - \Delta)$, we use $n-v$ examples to select a hypothesis, then use the remaining v examples to validate that hypothesis. For WAG to offer a stronger bound than SVOOSH, we must have $b(v, \epsilon - \Delta) < b(n-s, \epsilon)$. (For concentration inequalities $b()$ other than Hoeffding’s inequality, there exists a minimum s such that

$b(n-s, \epsilon) \leq m(n)b(n, \epsilon)$, and that s value plays the same role as our s . In luckiness frameworks and PAC-Bayesian frameworks, the s may depend on the selected hypothesis.)

The best case for WAG is $\Delta = 0$ – no disagreement between the holdout classifier and the one trained on all data. Even for this case, $b(v, \epsilon - \Delta) < b(n-s, \epsilon)$ requires $v > n-s$. So $n-s$ is a lower bound on the validation set size v if WAG is to be superior to SVOOSH. From Equality 2,

$$n-s = n - \frac{d}{2\epsilon^2} \ln n.$$

So

$$v > n - \frac{d}{2\epsilon^2} \ln n$$

is required for WAG to be superior. That leaves at most

$$w^* = \frac{d}{2\epsilon^2} \ln n$$

examples for training the holdout classifier in WAG.

In general, using fewer examples to train the holdout classifier results in a greater rate of disagreement Δ . So, for WAG to be more effective than SVOOSH, we can see that the hypothesis set dimension d must be large or the number of training examples n must be small, since $\ln n$ is a quickly-shrinking portion of n as n increases.

4 WAG vs. SVOOSH

To compare WAG to SVOOSH, let us examine the values of Δ needed for WAG to outperform SVOOSH. Set bound failure probability δ equal for both methods, let ϵ_W be the bound range for WAG, and let ϵ_V be the bound range for SVOOSH. Then

$$\epsilon_W = \Delta + \sqrt{\frac{\ln \frac{1}{\delta}}{2v}},$$

and

$$\epsilon_V = \sqrt{\frac{\ln \frac{1}{\delta} + \ln m(n)}{2n}}.$$

Let $m = n^d$. For $a > 1$, use $v = n/a$ examples for validation. Then $\epsilon_W < \epsilon_V$ (WAG superior to SVOOSH) requires

$$\Delta < \frac{1}{\sqrt{2n}} \left(\sqrt{\ln \frac{1}{\delta} + d \ln n} - \sqrt{a \ln \frac{1}{\delta}} \right).$$

Let Δ^* be the critical value for Δ :

$$\Delta^* = \frac{1}{\sqrt{2n}} \left(\sqrt{\ln \frac{1}{\delta} + d \ln n} - \sqrt{a \ln \frac{1}{\delta}} \right).$$

Figures 1, 2, and 3 show some values of Δ^* and ϵ_V . For all the plots, bound failure probability $\delta = 0.05$. Begin with Figure 1. It shows ϵ_V , the bound range for SVOOSH, and Δ^* , the critical rate of disagreement between holdout and full-data classifiers for WAG to outperform SVOOSH – below this rate of disagreement, WAG outperforms SVOOSH. For this plot, $d = 10$, meaning that the hypothesis class has dimension 10 ($m(n) = n^{10}$), and $a = 5$, meaning that one fifth of the examples are used for validation ($v = n/5$) and four fifths are used to train the holdout classifier. From 1000 to 10,000 training examples, Δ^* varies from about 0.10 to about 0.04, meaning that the rate of disagreement between a classifier trained on 4/5 of the examples and one trained on all examples must be less than about five to ten percent for WAG to be superior. This does not seem unreasonable.

The single-classifier bound range ϵ_W is the difference between the plotted lines: $\epsilon_V - \Delta^*$. The single-classifier bound range is about half the multiple-classifier bound range for all the numbers of

Critical Rate of Disagreement for $d = 10$ and $a = 5$

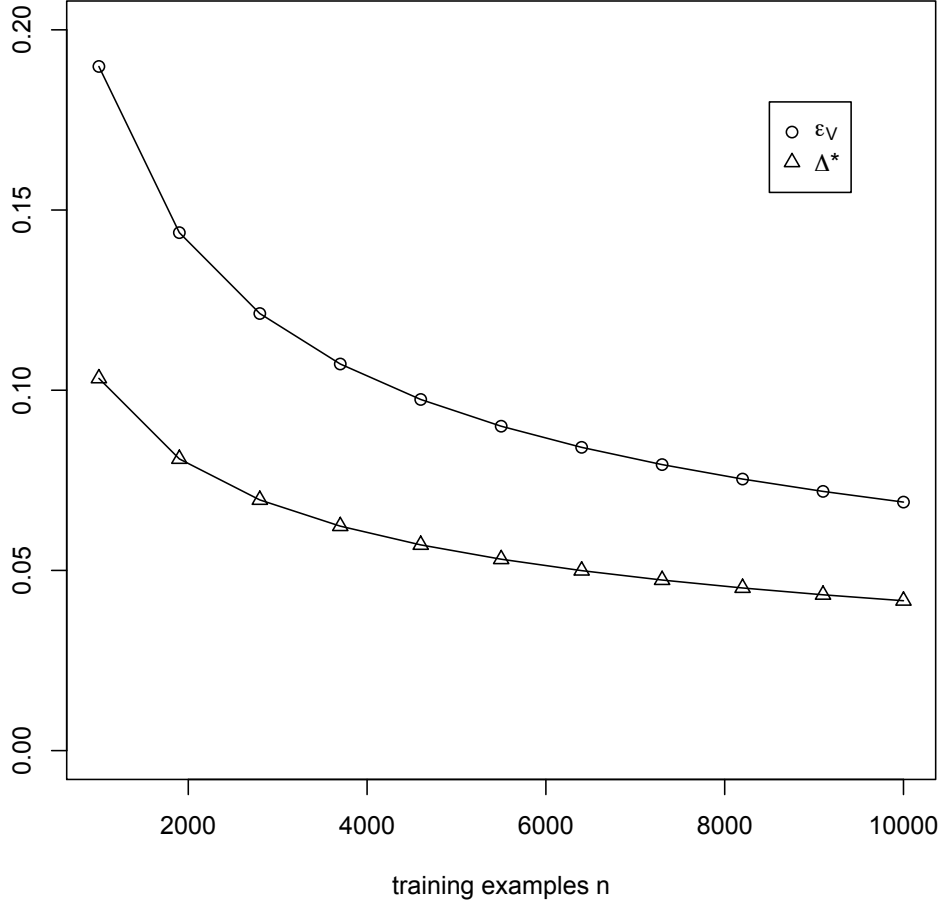


Figure 1: If there is less than about five to ten percent disagreement between holdout and full-data classifiers, WAG outperforms SVOOSH.

examples n shown in the graph. (This is multiple-classifier with n examples vs. single-classifier with $n/5$ examples.) This makes Δ^* about half the multiple-classifier bound range. So Δ^* is greatest for low numbers of training examples, where the multiple-classifier bound range is greatest.

Figure 2 shows Δ^* and ϵ_V for a less complex hypothesis class, with $d = 3$. With fewer hypotheses, the multiple-classifier bound range ϵ_V decreases compared to Figure 1. But the single-classifier bound range ϵ_W (not shown) remains the same. As a result, Δ^* , which is the difference between these bound ranges, gets squeezed. The figure shows Δ^* for $a = 3, 5$, and 10 , meaning that the validation sets are one third, one fifth, and one tenth of the training examples. For $a = 3$ and $a = 5$, Δ^* ranges from about one to five percent.

For $a = 10$, there are too few validation examples to perform effective enough single-classifier validation using $n/10$ examples to leave room for a reasonable Δ^* . For the lower values of n , the single-classifier bound based on one tenth of the examples is in fact worse than the multiple-classifier bound based on all examples. That makes Δ^* negative.

Figure 1 shows Δ^* and ϵ_V for $d = 100$. Due to the complexity of the hypothesis class, more training examples are required for SVOOSH to yield reasonable error bounds. The single-classifier bound ranges using validation data (ϵ_W , the gaps between ϵ_V and Δ^*) are several times smaller than the

Critical Rate of Disagreement for $d = 3$

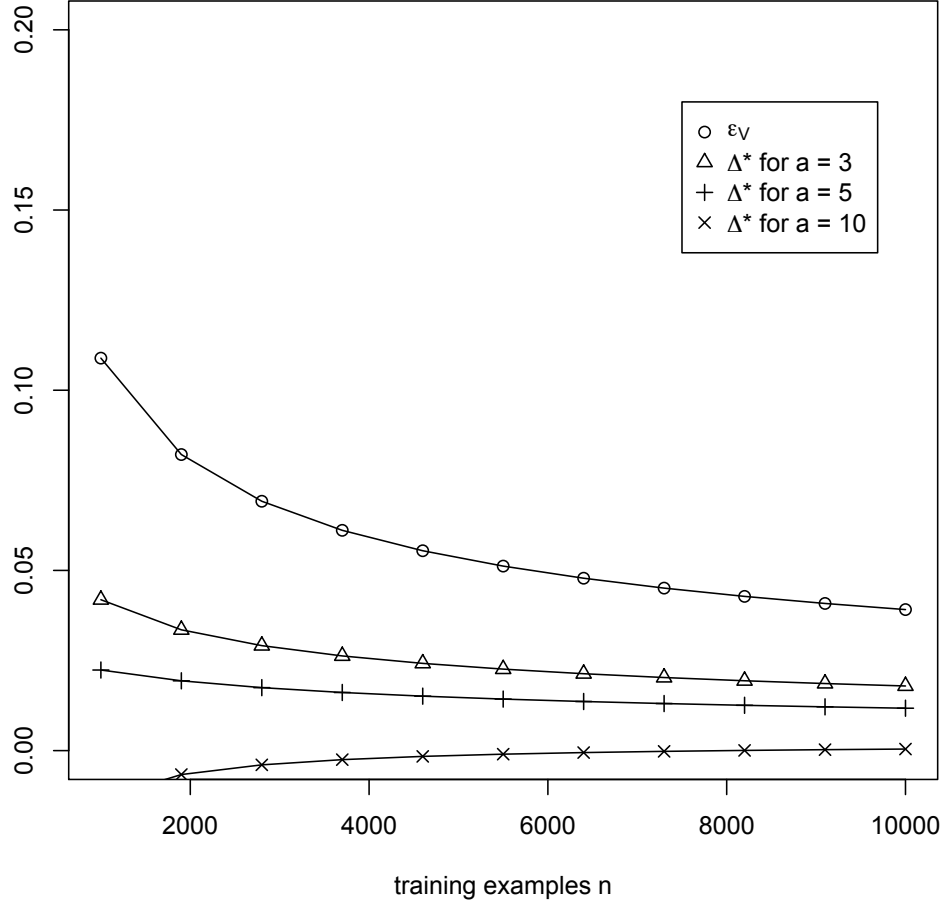


Figure 2: A lower-complexity hypothesis class leaves less room for WAG to outperform SVOOSH.

multiple-classifier bound ranges using all training data (ϵ_V). That leaves room for Δ^* values ranging from about seven to nineteen percent.

5 Discussion

We showed that WAG can outperform SVOOSH for validation of classifier training, using reasonable validation data set sizes, when the hypothesis set is sufficiently complex and there are few enough training examples, assuming reasonable rates of disagreement between holdout and full-data classifiers. After some consideration, WAG makes intuitive sense. If adding a few more training examples radically changes the resulting classifier, then it is hard to believe that the original (holdout) classifier would have generalized well. Similarly, it is easy to believe that adding a few more examples would result in yet again a very different classifier. In that case, it is hard to believe that the full-data classifier will generalize well.

How can we measure rate of disagreement Δ in a nontransductive setting? If there are unlabeled examples that are not used for learning, then we can use them to evaluate the rate of disagreement. If these are not the test examples, then the rate of disagreement over them is an empirical mean, and we can use concentration inequalities $b(t, \epsilon_T)$, where t is the number of unlabeled examples, to

Critical Rate of Disagreement for $d = 100$

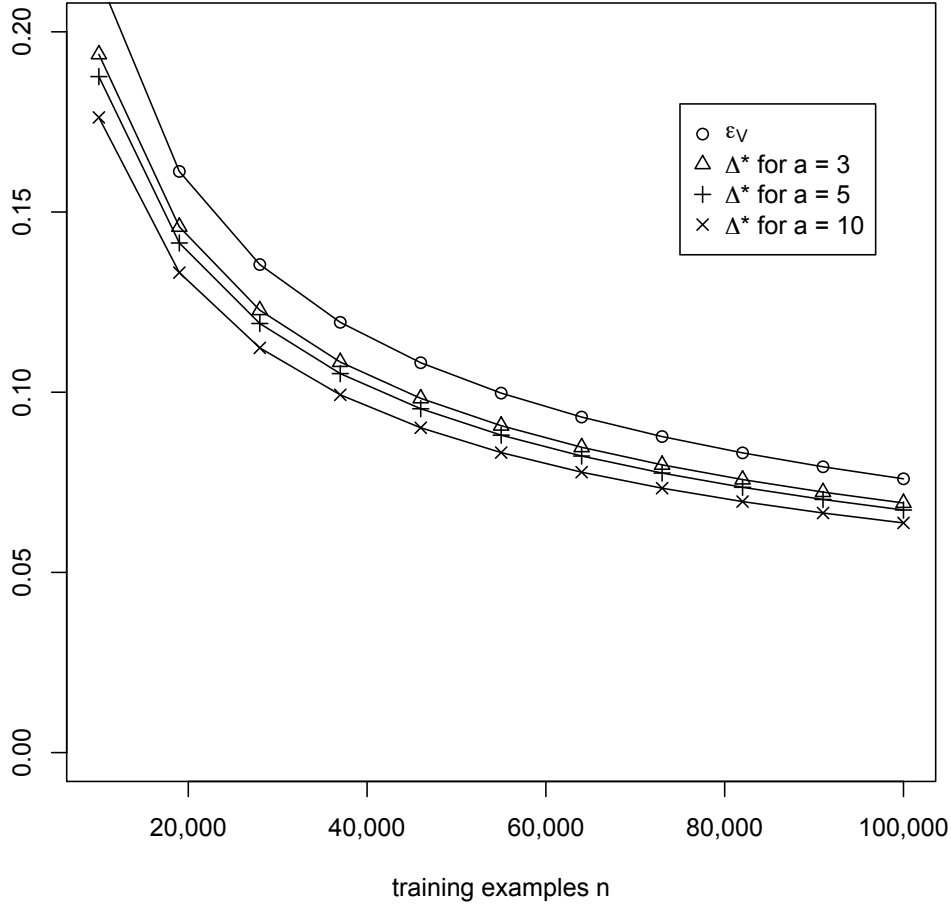


Figure 3: A higher-complexity hypothesis class makes more room for WAG to outperform SVOOSH.

bound the value of Δ over their generating distribution – presumably the same as the distribution of test examples. In this case, the WAG bound range is the sum of ϵ_W and ϵ_T . Fortunately, unlabeled examples are often easier to obtain than labeled ones, so ϵ_T can often be made small by taking a large number of unlabeled samples t .

It may be tempting to assert that the withheld validation examples can be used to simultaneously validate the holdout classifier and the rate of disagreement between the holdout classifier and the full-data classifier. However, the full-data classifier is not independent of the withheld data, since that data is part of its training set. An open question is whether there are conditions on training algorithms or hypothesis classes that allow this approach.

Recently, WAG has been used to validate network classifiers [13] and matching algorithms [14]. Central classifier bounds [15] are similar to the WAG approach analyzed here, but apply to validation of a classifier selected by early stopping. The method includes withholding a data set for validation and using it to validate the rate of disagreement between classifiers, but not using the validation data to train a full-data classifier. The general idea of withholding a data set for evaluation and then bounding the gap between the holdout classifier and a full-data classifier was first used for nearest neighbor classifiers, which have quite fractured decision boundaries, making it difficult to assess the size of their hypothesis sets. These methods [16, 17] do not empirically evaluate the

rate of disagreement between holdout and full-data classifiers; instead they take advantage of the locality of nearest neighbor classifiers and to prove upper bounds for the rates of disagreement based on geometry. As a result, they can be applied outside of transductive settings. More recent methods for nearest neighbors [18, 19] use empirical evaluation of rates of disagreement, including inclusion-exclusion techniques [20]. But they still rely on the locality of nearest neighbors classifiers to evaluate the gap between holdout and full-data classifiers, so it remains an open question whether such approaches can be applied to other types of classifiers.

References

- [1] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.
- [2] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.
- [3] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.
- [4] David McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.
- [5] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
- [6] K. Azuma. Weighted sums of certain dependent random variables. *Thoku Mathematical Journal*, 19(3):357–367, 1967.
- [7] C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics, London Math. Soc. Lecture Notes*, 141:148–188, 1989.
- [8] J. Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.
- [9] P. G. Hoel. *Introduction to Mathematical Statistics*. Wiley, 1954.
- [10] J.-Y. Audibert, R. Munos, and Csaba Szepesvari. Variance estimates and exploration function in multi-armed bandit. *CERTIS Research Report 07-31*, 2007.
- [11] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *COLT*, 2009.
- [12] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities – A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.
- [13] E. Bax, J. Li, A. Sonmez, and Z. Cataltepe. Validating collective classification using cohorts. *NIPS Workshop on Frontiers of Network Analysis: Methods, Models, and Applications*, 2013.
- [14] Ya Le, Eric Bax, Nicola Barbieri, David Garcia Soriano, Jitesh Mehta, and James Li. Validation of network reconciliation. <http://arxiv.org/abs/1411.0023>, 2015.
- [15] Eric Bax, Zehra Cataltepe, and Joseph Sill. The central classifier bound – a new error bound for the classifier chosen by early stopping. *IEEE PACRIM Conference, Victoria, Canada*, pages 811–814, 1997.
- [16] L. Devroye and T. Wagner. Distribution-free inequalities for the deleted and holdout estimates. *IEEE Transactions on Information Theory*, 25:202–207, 1979.
- [17] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.
- [18] E. Bax. Validation of nearest neighbor classifiers. *IEEE Transactions on Information Theory*, 46(7):2746–2752, 2000.
- [19] E. Bax. Validation of k -nearest neighbor classifiers. *IEEE Transactions on Information Theory*, 58(5):3225–3234, 2012.
- [20] Eric Bax, Lingjie Weng, and Xu Tian. Validation of k -nearest neighbor classifiers using inclusion and exclusion. <http://arxiv.org/abs/1410.2500>, 2015.